Procesamiento del Lenguaje Natural, Revista nº 44, marzo de 2010, pp 159-160 recibido 19-01-10 revisado 16-02-10 aceptado 08-03-10

Hierarchical self-refining consensus architectures and soft consensus functions for robust multimedia clustering

Arquitecturas de consenso jerárquicas auto-refinables y funciones difusas de consenso para agrupamiento robusto de datos multimedia

Xavier Sevillano
GTM – Grup de Recerca en Tecnologies Mèdia
La Salle. Universitat Ramon Llull
C/Quatre Camins, 2. 08022 Barcelona
xavis@salle.url.edu

Resumen: Tesis doctoral en Tecnologías de la Información y las Comunicaciones y su Gestión. El acto de defensa de tesis tuvo lugar en Barcelona en Junio de 2009 ante el tribunal formado por los doctores Paolo Rosso (Univ. Politécnica de Valencia), Aristides Gionis (Yahoo! Research), Juan José Rodríguez (Univ. de Burgos), Jordi Turmo (Univ. Politécnica de Catalunya) y Ester Bernadó (La Salle - Univ. Ramon Lull). La calificación obtenida fue Sobresaliente Cum Laude.
Palabras clave: Agrupamiento robusto, consenso, agrupamiento difuso, multimedia

Abstract: PhD thesis in Information and Communication Technologies and their Management. The author was examined in June 2009 in Barcelona by the following committee: Paolo Rosso (Univ. Politécnica de Valencia), Aristides Gionis (Yahoo! Research), Juan José Rodríguez (Univ. de Burgos), Jordi Turmo (Univ. Politécnica de Catalunya) and Ester Bernadó (La Salle - Univ. Ramon Lull). The grade obtained was Summa Cum Laude.
Keywords: Robust clustering, consensus, fuzzy clustering, multimedia

1 Introduction

The robust design of clustering systems is a very relevant and challenging issue. This is due to the unsupervised nature of the clustering problem, which makes it difficult (if not impossible) to select *a priori* the configuration of the clustering system that gives rise to the most meaningful partition of a data collection. Furthermore, given the myriad of options –e.g. clustering algorithms, data representations, etc.– available to the clustering practitioner, such important decision is often made with a high degree of uncertainty (unless domain knowledge is available, which is not always the case).

For this reason, our approach to robust clustering intentionally reduces user decision making as much as possible: the clustering practitioner is encouraged to use *and* combine all the clustering configurations at hand, compiling the resulting clusterings into a cluster ensemble, upon which a consensus clustering is derived. The more similar this consensus clustering is to the highest quality clustering contained in the cluster ensemble, the greater the robustness to the indeterminacies inherent to clustering.

This PhD thesis is focused on the problem of robust clustering based on cluster ensembles, with a specific focus on the increasingly interesting application of multimedia data clustering and a view on its generalization in fuzzy clustering scenarios.

More specifically, the main goal of this research work is to derive *high quality* consolidated clusterings upon cluster ensembles in a *computationally efficient* manner. This latter issue is especially relevant, as our particular approach to robust clustering indirectly entails the creation of *large* cluster ensembles, a fact that dramatically increases execution time and memory usage of consensus clustering algorithms.

Furthermore, our proposals find a natural field of application in *multimedia data clustering*, as the existence of multiple data

ISSN 1135-5948

modalities poses additional indeterminacies that challenge the obtention of robust clustering results.

Finally, in order to generalize our approach to robust clustering based on cluster ensembles, we propose several voting based consensus functions for deriving *fuzzy consensus partitions* by combining of the outcomes of multiple *fuzzy clustering systems*.

2 Thesis overview

The PhD thesis is organized as follows: the first two chapters provide an introduction to the central matter of the thesis. In particular, chapter 1 presents an overview of the clustering problem, and chapter 2 reviews related work in the field of consensus clustering.

Chapter 3 introduces hierarchical consensus architectures, our proposal for the computationally efficient derivation of consensus clusterings based on the application of the divide-and-conquer strategy on cluster ensembles.

In chapter 4, we present self-refining consensus clustering, a fully unsupervised methodology for obtaining high quality consensus partitions based on excluding low quality components of the cluster ensemble from the consensus process.

Chapter 5 shows how our proposals for robust clustering based on cluster ensembles naturally allow the simultaneous use of early and late multimodal fusion techniques, thus constituting a highly generic approach to the problem of multimedia data clustering.

In chapter 6, we present voting based consensus functions for combining the outputs of multiple fuzzy unsupervised classifiers, a first step for porting our previous proposals to the more generic framework of fuzzy clustering.

Finally, in chapter 7 we discuss the main conclusions of our work, outlining several future research lines of interest that stem from the investigation presented in this thesis.

3 Thesis contributions

The major contributions of this PhD thesis are the following:

- hierarchical consensus architectures constitute a highly parallelizable strategy for the fast derivation of consensus clusterings upon cluster ensembles. Besides defining random and deterministic hierarchical consensus architectures, we also design simple tools that allow to predict accurately the most computationally efficient hierarchical consensus architecture for a given consensus clustering problem.
- the self-refining consensus clustering proposal is capable of generating a bunch of self-refined consensus clusterings upon a previously obtained consensus partition, some of which are a largely improved version of the latter. We complement our approach by introducing a blind strategy for selecting the optimal self-refined consensus clustering among them.
- multimedia clustering based on cluster ensembles starts by creating clusterings upon each separate modality *and* on feature level fused modalities, and after compiling them all into a multimodal cluster ensemble, a consensus clustering is created upon it. By doing so, we take advantage of modality fusion both at feature level (early fusion) and at decision level (late fusion).
- voting based consensus functions for soft cluster ensembles allow to obtain fuzzy consensus partitions. Among our proposals, the BordaConsensus and CondorcetConsensus algorithms are pioneer consensus functions based on positional voting.

Throughout this thesis, we have given great importance to the experimental evaluation of all our proposals, using a total of 16 unimodal and multimodal publicly available data collections. The results of these experiments have confirmed that, in the quest for the design of robust multimedia clustering systems in which user decision making is minimized, our cluster ensembles-based proposal succeeds in the computationally efficient derivation of high quality consensus partitions. In the future, we plan to set the user free from the obligation to select how many clusters the objects are to be grouped in, using this element as an additional factor of diversity at the time of creating the cluster ensemble.